IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Date of current version 6 September 2022.

Digital Object Identifier 10.1109/TETC.2022.3195414

Guest Editorial: Special Section on Emerging and Impacting Trends on Computer Arithmetic

STUART OBERMAN, (Senior Member, IEEE), LEONEL SOUSA, (Senior Member, IEEE), BOGDAN PASCA, (Member, IEEE), AND ALBERTO NANNARELLI, (Senior Member, IEEE)

The Computer Arithmetic field encompasses the definition and standardization of arithmetic systems for computers. It also deals with issues pertaining to hardware and software implementations, testing, and verification. Researchers and practitioners of this field also work on challenges associated with using Computer Arithmetic to perform scientific and engineering calculations. As such, Computer Arithmetic can be regarded as a truly multidisciplinary field, which builds upon mathematics, computer science and electrical engineering. Thus, the range of topics addressed by Computer Arithmetic is generally very broad, spanning from highly theoretical to extremely practical contributions. Computer Arithmetic has been an active research field since the advent of computers, and it is progressively evolving following continuous advancements in technology.

To acknowledge the above richness, this Special Section called for submissions in a number of emerging domains of Computer Arithmetic, including innovative number systems, floating point units and algorithms, high-level language and compiler impact on arithmetic systems, approaches to test, verification, formal proof, computer aided design (CAD) automation and fault/error tolerance for arithmetic architectures, paradigms for FPGAs or configurable logic, inexact and stochastic arithmetic, as well as efficient, low-power and novel implementations. Submissions related to new arithmetic paradigms and architectures for specific application domains such as cryptography, security, neural networks, deep learning, signal processing, computer graphics, multimedia, computer vision, distributed and parallel computing (e.g., HPC), and finance, among others, were also solicited.

The Special Section was launched jointly with the 29th IEEE International Symposium on Computer Arithmetic (ARITH 2022). Since 1969, ARITH has been the premier international event for Computer Arithmetic research. For the second time, ARITH featured two categories of submissions: Journal Papers (JPs), which were scheduled to appear in this Special Section according to the publication framework known as J1C2 (Journal 1st, Conference 2nd), and Conference Papers (CPs). The two categories had separate submission, review and publication processes, and all accepted JPs also included an oral presentation in a regular session at ARITH 2022, like CP submissions. This Special Section received 28 submissions, which were prescreened and reviewed by experts in the field. The following seven papers were ultimately accepted for publication.

The article "PERCIVAL: Open-Source Posit RISC-V Core With Quire Capability" by David Mallasén Quintana, Raul Murillo, Alberto A. Del Barrio, Guillermo Botella, Luis Piñuel, and Manuel Prieto-Matias explores the use of dedicated hardware units operating on the Posit representation as hardware extensions of a RISC-V core. The proposed hardware units are combined with a complete implementation of the LLVM XPosit extension into an Open-Source Tool – Percival. The authors assess the accuracy of the Posit-based implementation against an IEEE-754 single-precision implementation for various input ranges in a GEMM application.

The article "Bounding the Round-Off Error of the Upwind Scheme for Advection" by Louise Ben Salem-Knapp, Sylvie Boldo, and William Weens tackles the challenges of bounding round-off errors for a complex system from hydrodynamics. More precisely – the paper deals with the propagation of the rounding error when computing the solution to a specific partial differential equation using a numerical scheme in floating-point arithmetic. The authors prove that the global rounding error grows linearly with respect to the number of timesteps and, consequently, are able to provide tight error bounds.

The article "PMNS for Efficient Arithmetic and Small Memory Cost," by Fangan Yssouf Dosso, Jean-Marc Robert, and Pascal Véron is on the Polynomial Modular Number System (PMNS), for speeding up arithmetic operations modulo a prime number p. PMNS has received a lot of interest recently. The authors provide a method to generate multiple PMNS for a given prime p, generalizing a previous approach proposed for a particular class of PMNS (AMNS). Experimental results show that the proposed PMNS are as efficient as the polynomials for AMNS. Moreover, it appears that PMNS with the Montgomery internal reduction is well suited for vectorization, considering the significant speedup achieved when the SIMD AVX512 instruction set is applied.

The article "An Alternative Approach to Polynomial Modular Number System Internal Reduction," by Nicolas Méloni improves the internal reduction operation on PMNS. Alternative to using Montgomery's modular multiplication, this article proposes new techniques based on the Babai's Closest

Vector algorithms to perform the internal reduction. It significantly reduces the number of additions needed to perform this operation. A comprehensive experimental analysis shows that these algorithms, in particular one of them, is faster in practice than the state-of-the-art.

In the article "Generating Very Large RNS Bases," Jean Claude Bajard, Kazuhide Fukushima, Thomas Plantard, and Arnaud Sipasseuth address the question of finding the existence of large RNS bases for some specific uses, such as in the context of cryptographic engineering. They extend a work presented in ARITH 2021 that introduced a first approach of filtering selection of RNS bases belonging to a given specific interval. Limiting the set of possible pairwise co-primes, new filtering algorithms with selection functions are proposed, for efficiently providing the largest possible bases of co-primes moduli.

In their article "A BF16 FMA is All You Need for DNN Training," John Osorio, Adrià Armejach, Eric Petit, Greg Henry, and Marc Casas propose an approach to training complex Deep Neural Networks (DNNs) entirely in the BrainFloat16 (BF16) floating point format. Using a narrower floating-point format such as BF16 can accelerate the training of DNNs, but such training is typically combined with other higher precision formats as needed to achieve high accuracy. The authors propose a new class of BF16 FMA operators that when used in DNN training can enable the same accuracy as training runs using single precision floating point (FP32), while only using the BF16 format.

Lastly, the focus of the article "Approximate Recursive Multipliers Using Low Power Building Blocks" by Efstratios Zacharelos, Italo Nunziata, Gerardo Saggese, Antonio G. M. Strollo, and Ettore Napoli, is on approximate computing, frequently used in error tolerant applications. Approximate computing circuits can achieve higher performance and lower power by allowing the possibility of inaccurate results, rather than guaranteeing a correct outcome. This paper explores designing multipliers, typically high-power circuits, recursively using lower-power approximate building blocks. In this way, their technique enables the construction of lower power multipliers while maintaining competitive error.

The topics tackled by these papers clearly show how rich and diverse Computer Arithmetic can be, hopefully indicating to interested readers possible directions for further research in this field.

On behalf of every reader of this Special Section, the Guest Editors would like to thank all the authors who submitted their papers and worked hard to respond to Reviewers' requests in due time, all the anonymous reviewers who participated in the review process providing helpful suggestions, as well as the Editor in Chief and the entire staff of IEEE Transactions on Emerging Topics in Computing who oversaw the whole process.

STUART OBERMAN (Senior Member, IEEE) received the BS degree in electrical engineering from the University of Iowa, and the MS and PhD degrees in electrical engineering from Stanford University, where he performed research with the Stanford Architecture and Arithmetic Group. He is vice president of GPU Engineering with NVIDIA. Since 2002, he has contributed to the design and verification of ten GPU architectures. He currently directs multiple GPU design and verification teams. He has co-authored one book and more than 20 technical papers. He holds more than 55 granted US patents. He is a senior member of the IEEE Computer Society.

LEONEL SOUSA (Senior Member, IEEE) received the PhD degree in electrical and computer engineering from the Instituto Superior Tecnico (IST), Universidade de Lisboa (UL), Lisbon, Portugal, in 1996, where he is currently full professor. His research interests include VLSI architectures, computer arithmetic, computer architectures and parallel processing. He has contributed to more than 300 papers in journals and international conferences. He has edited five special issues of international journals, and he is currently associate editor of the *IEEE Transactions on Computers*, and senior editor of the *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*. He has been a distinguished lecturer with the DLP IEEE CAS society program (2017-2019), a distinguished visitor in the DVP from the IEEE CS (2018-2020), and an ACM Distinguished Speaker. He is a fellow of the IET, distinguished scientist of the ACM and distinguished contributor of the IEEE Computer Society.

BOGDAN PASCA (Member, IEEE) received the MSc and PhD degrees from the École Normale Supérieure de Lyon, in 2008 and 2011, respectively. He is currently with the Intel Programmable Solutions Group. His research interests include computed arithmetic and reconfigurable computing. Over the last decade with Altera, and then Intel, he has been architecting and implementing an extensive library of arithmetic cores - fixed and floating-point - which are currently used throughout the majority of Intel FPGA products. His work received the Michal Servit Memorial Award at the 22nd International Conference on Field Programmable Logic and Applications (FPL) and the Best Paper Award at the 31st IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP) in 2020.

ALBERTO NANNARELLI (Senior Member, IEEE) graduated in electrical engineering from the University of Rome "La Sapienza," Italy, in 1988, and received the MS and PhD degrees in electrical and computer engineering from the University of California at Irvine, in 1995 and 1999, respectively. He is an Associate Professor with the Technical University of Denmark, Lyngby, Denmark. He worked for SGS-Thomson Microelectronics and for Ericsson Telecom as a design engineer and for Rockwell Semiconductor Systems as a summer intern. From 1999 to 2003, he was with the Department of Electronic Engineering, University of Rome "Tor Vergata," Italy, as a postdoctoral researcher. His research interests include computer arithmetic, computer architecture, and VLSI design. He is a senior member of the IEEE Computer Society.